THỐNG KÊ TOÁN HỌC

HỒI QUY và TƯƠNG QUAN

Khái niệm cơ bản

Ví dụ

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287	2015	?
1970	3712	1995	5700	2020	?

TABLE 11.1: *Population of the world, 1950–2020.*

Ví dụ

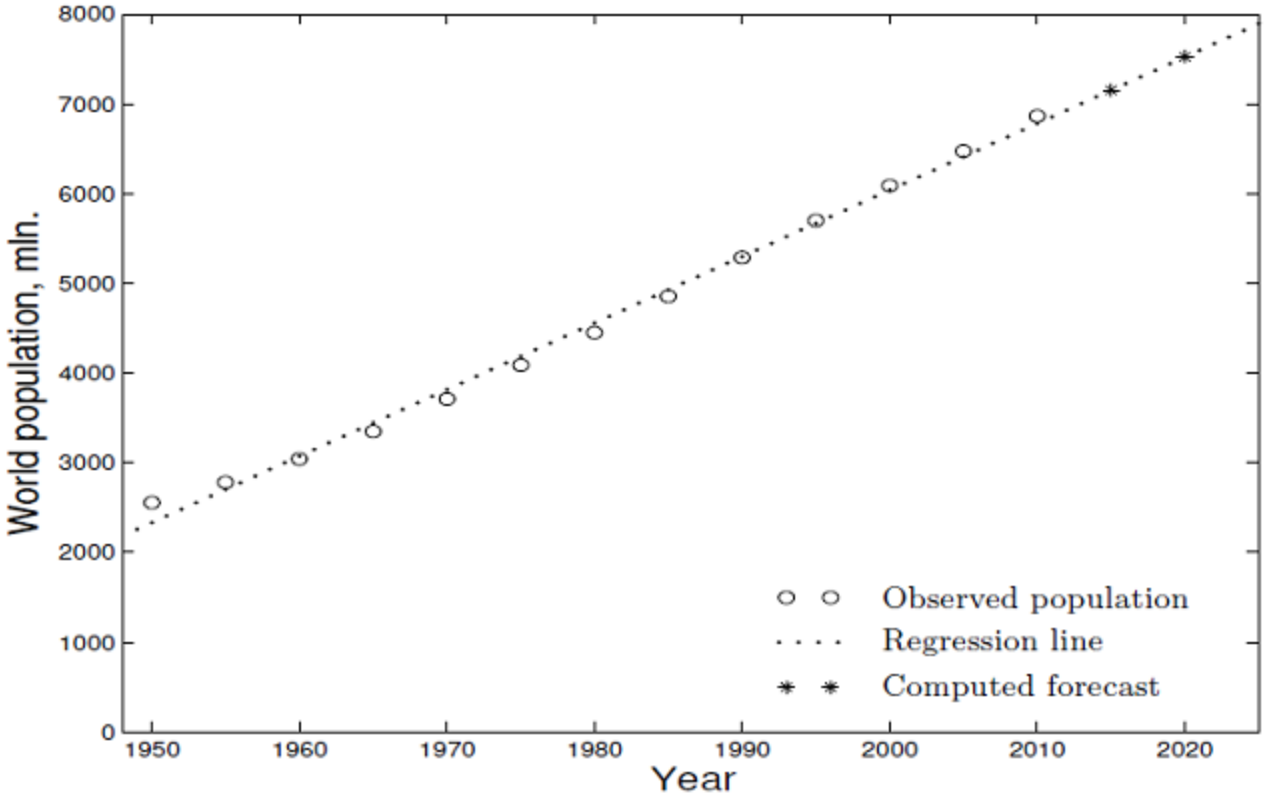
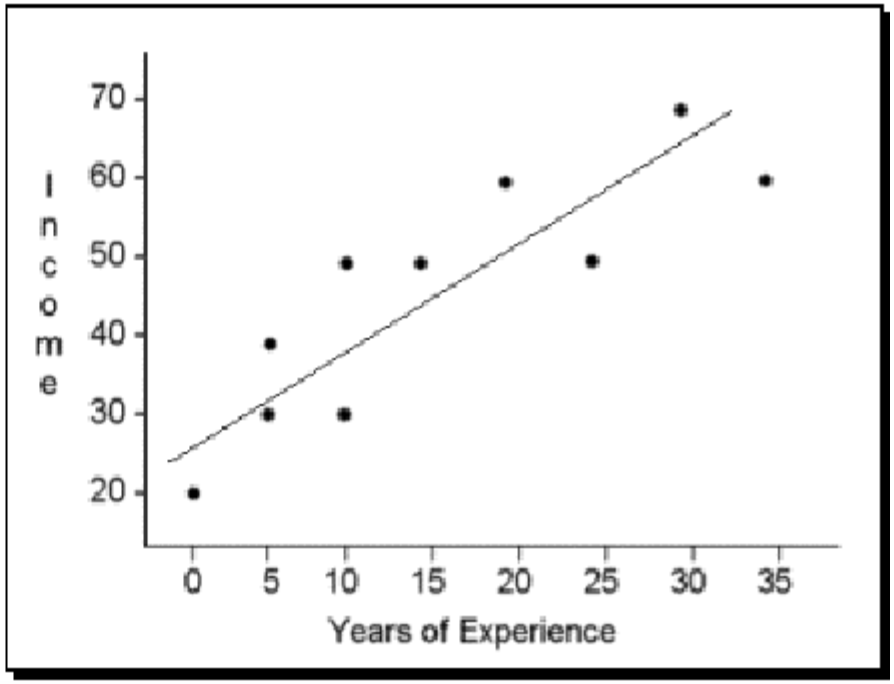


FIGURE 11.1: *World population in 1950–2010 and its regression forecast for 2015 and 2020.*

Biểu đồ phân tán

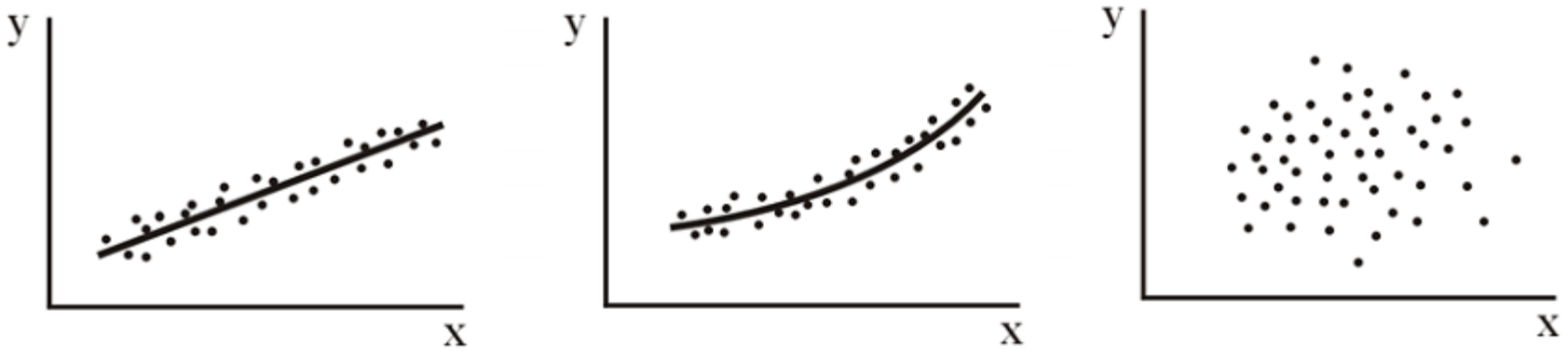
- Khảo sát sự tương quan giữa X và Y.
  - Vd: thu nhập và số năm kinh nghiệm.
- Xét 2 đại lượng ngẫu nhiên đồng thời (X, Y) và tập n cặp giá trị cụ thể $(x_1, y_1), \dots, (x_n, y_n)$
- Các giá trị (x_i, y_i) có được từ việc khảo sát số liệu, do đó đc gọi là dữ liệu thực nghiệm.



Biểu đồ phân tán: tập hợp các điểm (x_i, y_i) trên mặt phẳng tọa độ.

Đường cong phù hợp

- Đường cong phù hợp: là một đường cong ***xấp xỉ tốt*** *(ít sai lệch nhất)* với dữ liệu đã cho.
  - Nếu đường cong phù hợp là một đường thẳng: ta có một quan hệ tuyến tính giữa các đại lượng.
  - Nếu đường cong phù hợp không là một đường thẳng: ta có một quan hệ phi tuyến giữa các đại lượng.



Vấn đề đặt ra

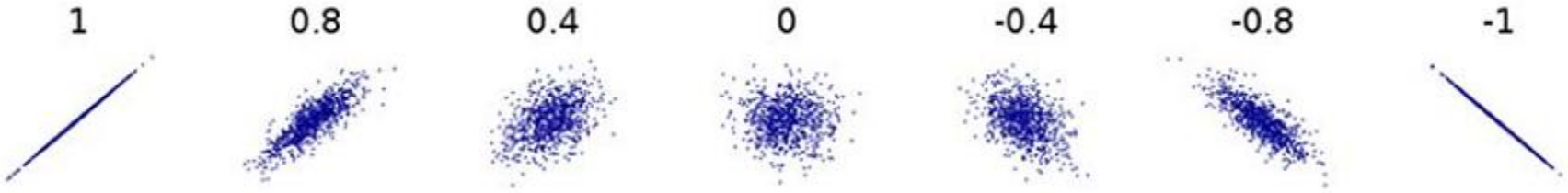
- Có một sự **tương quan** (tuyến tính hoặc phi tuyến) nào giữa các đại lượng hay không?
  - *Ví dụ:* chiều dài và cân nặng của một con cá mập có mối liên hệ gì với nhau? Tuyến tính hay phi tuyến?
- Nếu có, thì làm sao để viết được một phương trình **hồi quy** biểu thị mối quan hệ này?
  - Phương trình của cân nặng ***theo*** chiều dài?
  - Dự đoán cân nặng của một con cá mập khi biết chiều dài của nó?

Hệ số tương quan

Hệ số tương quan mẫu

- Hệ số tương quan mẫu r: dùng để đo sự ***tương quan tuyến tính*** giữa 2 đại lượng X và Y.

- r là một **con số**, không có đơn vị, nằm trong khoảng từ -1 đến 1.



- *Nhận xét:* |r| càng gần 1, càng có sự tương quan tuyến tính giữa 2 đại lượng.

Hệ số tương quan mẫu

- Công thức

$$r = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - (\overline{x})^2\right)\left(\overline{y^2} - (\overline{y})^2\right)}}$$

- $|r| > 0{,}9$: có sự tương quan tuyến tính mạnh giữa X, Y
  - $r > 0$: tương quan tuyến tính **thuận** (X, Y đồng biến)
  - $r < 0$: tương quan tuyến tính **nghịch** (X, Y nghịch biến)
- $|r| < 0{,}9$: tương quan tuyến tính giữa X và Y yếu.
- |r| càng gần 1 thì sự tương quan tuyến tính càng mạnh

Ví dụ

Điểm kiểm tra môn XSTK và môn Giải tích của 10 sinh viên được chọn ngẫu nhiên từ lớp có rất nhiều sinh viên:

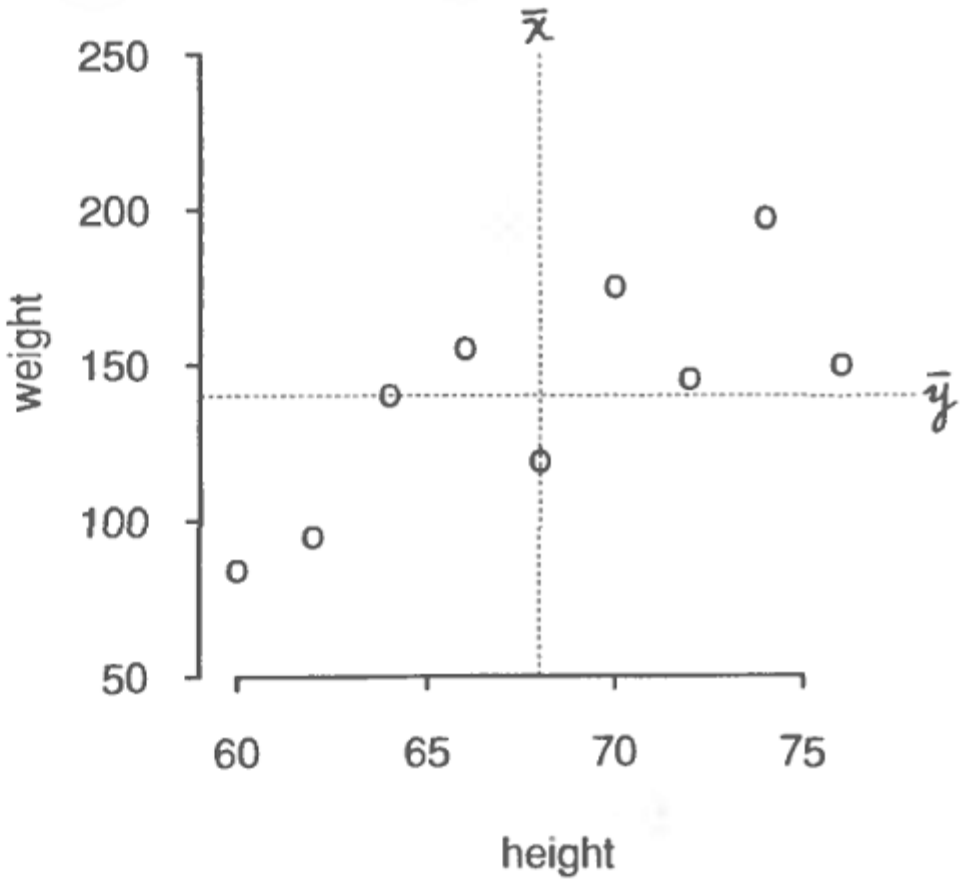
XSTK	73	80	93	65	87	71	98	68	84	70
Giải tích	82	79	86	72	91	80	97	72	89	74

Tính hệ số tương quan của điểm môn Giải tích và môn XSTK.

Bài tập

Tìm hệ số tương quan tuyến tính giữa chiều cao và cân nặng.

HEIGHT	WEIGHT
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



Bài toán hồi quy

Hồi quy

- Mục đích: **dự đoán đại lượng này** (*đại lượngphụ thuộc*) **từ các đại lượng khác** (*các đại lượngđộc lập*).
  - Vd: biết chiều dài của một con cá mập → cân nặng, tuổi, chiều rộng hàm cá mập?
- Tiến trình ước lượng này được gọi là tiến trình hồi quy.
- Nếu Y được ước lượng từ X bằng một biểu thức nào đó thì ta gọi **biểu thức này** là phương trình hồi quy *của Y theo X*.
- **Đường cong tương ứng**được gọi là đường cong hồi quy*của Y theo X.*
  - **Đường thẳng tương ứng** được gọi là đường thẳng hồi quy*của Y theo X.*

Hồi quy tuyến tính

Đường thẳng bình phương bé nhất

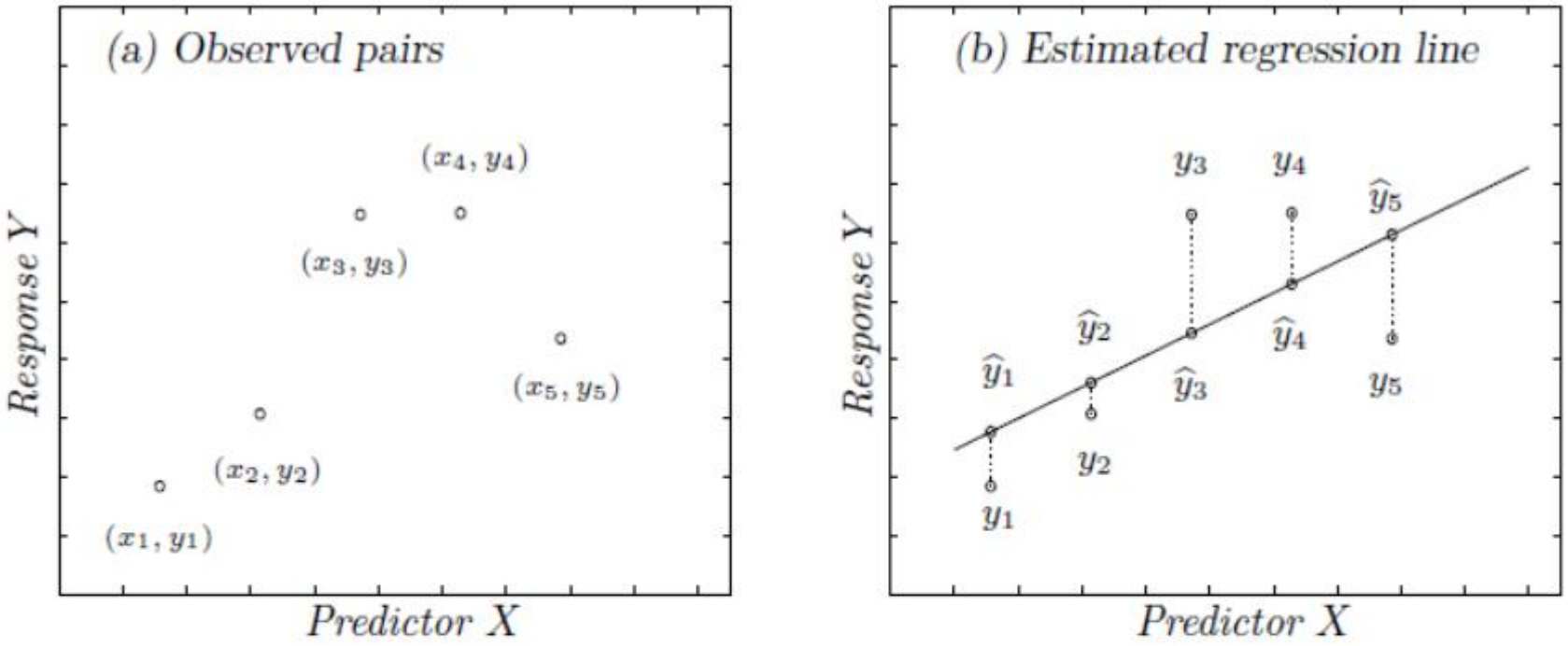


FIGURE 11.3: *Least squares estimation of the regression line.*

Giả sử PT đường thẳng là Y = A + BX

Tổng bình phương độ lệch: $\sum_{i=1}^{n}(y_i - (A + Bx_i))^2 = g(A, B)$

Phương trình hồi quy tuyến tính

Phương trình hồi qui tuyến tính của Y theo X:

$$Y = A + B\,X$$

Trong đó $B = \dfrac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - (\bar{x})^2}$ và $A = \bar{y} - B.\bar{x}$

$\bar{x} = \dfrac{\sum x_k}{n}$	$\bar{y} = \dfrac{\sum y_k}{n}$	$\overline{xy} = \dfrac{\sum x_k y_k}{n}$	$\overline{x^2} = \dfrac{\sum x_k^2}{n}$	$\overline{y^2} = \dfrac{\sum y_k^2}{n}$

Sử dụng máy tính

Ví dụ Bài toán cho dạng cặp (x_i, y_i) như sau:

X	20	52	30	57	28	43	57	63	40	49
Y	1,9	4	2,6	4,5	2,9	3,8	4,1	4,6	3,2	4

Tìm hệ số tương quan r_{xy}, đường hồi qui mẫu $y = a + bx$.

a. Máy fx-570VN Plus

- Bước 1: SHIFT; MODE; ↓; chọn 4 (Stat); chọn 2 (Off)
- Bước 2: MODE; chọn 3 (Stat); chọn 2 (A+Bx)
  - Nhập giá trị của X 20= 52= $\cdots$
  - Nhập giá trị của Y 1.9= 4= $\cdots$
- Bước 3: Xuất kết quả On; SHIFT; chọn 1 (Stat); chọn 5 (Reg); 1(A =a); 2(B=b); 3(r=r_{xy}).

Kết quả $r_{xy} = 0,9729; y = 0,9311 + 0,0599x$.

Sử dụng máy tính

Ví dụ Bài toán cho dạng cặp (x_i, y_i) như sau:

X	20	52	30	57	28	43	57	63	40	49
Y	1,9	4	2,6	4,5	2,9	3,8	4,1	4,6	3,2	4

Tìm hệ số tương quan r_{xy}, đường hồi qui mẫu $y = a + bx$.

b. Máy fx-580VN X

- Bước 1: SHIFT; MENU; ↓; chọn 3 (Stat); chọn 2 (Off)
- Bước 2: MENU; chọn 6 (Stat); chọn 2 (y=ax+b)
  Nhập giá trị của x 20= 52= …
  Nhập giá trị của y 1.9= 4=…
- Bước 3: Xuất kết quả OPTN; chọn 4

Kết quả $r_{xy} = 0,9729; y = 0,9311 + 0,0599x$.

Ví dụ

- Cho 2 ĐLNN (X, Y) với các giá trị tương ứng trên một mẫu như sau:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

- Tìm phương trình đường thẳng bình phương bé nhất với Y là biến độc lập và X là biến phụ thuộc.
- Tìm phương trình đường thẳng bình phương bé nhất với X là biến độc lập và Y là biến phụ thuộc.
- Hãy ước lượng giá trị của y khi $x_0 = 12$ và giá trị của x khi $y_0 = 3$.

Lưu ý

- Đường thẳng hồi quy theo phương pháp bình phương bé nhất luôn đi qua điểm $(\bar{x}, \bar{y})$.
- Khi tính toán cần xác định rõ biến độc lập và biến phụ thuộc:

PT hồi qui của Y theo X:

$$Y = A + B X$$

PT hồi qui của X theo Y:

$$X = C + D Y$$

Ví dụ

Điểm kiểm tra môn XSTK và môn Giải tích của 10 sinh viên được chọn ngẫu nhiên từ lớp có rất nhiều sinh viên:

XSTK	73	80	93	65	87	71	98	68	84	70
Giải tích	82	79	86	72	91	80	97	72	89	74

1. Tìm phương trình hồi quy tuyến tính của điểm môn Giải tích theo môn XSTK.
2. Nếu một sinh viên có 85 điểm môn Giải tích thì hy vọng sinh viên đó có bao nhiêu điểm môn XSTK?

Bài tập

Thu thập số liệu về số giờ tự học trong 1 tuần và số tín chỉ theo học của một sinh viên.

Số giờ tự học	20	25	24	30	32	45
Số tín chỉ	12	13	12	15	14	16

1. Tìm hệ số tương quan giữa 2 ĐLNN này.
2. Tìm hàm hồi quy tuyến tính của số giờ tự học theo số tín chỉ.
3. Hãy dự đoán số giờ tự học trong 1 tuần của sinh viên khi theo học 17 tín chỉ.
4. Tìm hàm hồi quy tuyến tính của số tín chỉ theo số giờ tự học.
5. Nếu một sinh viên tự học 26 giờ/tuần thì dự đoán sinh viên này đã đăng ký bao nhiêu tín chỉ?